GDPval: AIモデルの経済的価値評価に関する包括的調査報告書

Gemini 3 pro

1. エグゼクティブサマリー

人工知能（AI）技術の急速な進展に伴い、その能力を測定するための評価指標（ベンチマーク）の在り方が根本的な転換期を迎えている。本報告書は、OpenAIによって開発された新しい評価フレームワークである「GDPval（Gross Domestic Product valuable tasks）」について、その構造的特徴、初期評価結果、方法論的課題、および経済的・社会的影響を包括的に調査・分析したものである。

GDPvalは、従来の「MMLU」や「GSM8K」といった学術的な推論能力テストとは一線を画し、米国国内総生産（GDP）への貢献度が高い9つの主要セクター、44の職業から抽出された「経済的に価値のあるタスク」をAIに遂行させ、その成果物を人間の専門家が評価するという画期的なアプローチを採用している[1]。最新のGPT-5.2 Thinkingモデルは、この指標において人間の専門家に対し70.9%の勝率（引き分け含む）を記録し、特定の知識労働領域においてAIが人間と同等以上の成果を出せる可能性を示唆した[4]。

しかし、本調査の結果、GDPvalには「現実世界の複雑性」を捨象しているという重大な批判が存在することが明らかになった。特に、より実践的で曖昧性の高いタスクを扱う「Remote Labor Index（RLI）」におけるAIの成功率がわずか2.5%にとどまっている事実は、GDPvalが示す「人間並み」という評価が、明確に定義された限定的な条件下でのみ成立することを示唆している[6]。また、評価プロセスにおける人間とAIの一致率が66%にとどまるという「評価のボトルネック」は、AIの社会実装における品質保証の難しさを浮き彫りにしている[8]。

本報告書では、GDPvalの技術的詳細と評価結果を詳述するとともに、それが企業のROI（投資対効果）予測や労働市場の再編に与える影響、そして「タスクの自動化」と「職業の代替」の間に横たわるギャップについて、多角的な視点から論じる。

---

2. 序論：AI評価のパラダイムシフト

2.1 学術的知能から経済的有用性へ

長年、大規模言語モデル（LLM）の評価は、主に認知科学や言語学に由来する学術的なテストによって行われてきた。MMLU（Massive Multitask Language Understanding）による多肢選択問題、GSM8Kによる数学的推論、HumanEvalによるPythonコード生成などは、モデルの「知能」の代理指標として機能してきた。しかし、これらのベンチマークで高いスコアを記録したモデルであっても、実際のビジネス現場では「役に立たない」あるいは「期待外れ」であるという事例が後を絶たなかった。これは、学術的な問題解決能力と、現実の業務で求められる成果物（Deliverables）作成能力との間

に、非自明な乖離が存在することを示している[2]。

この背景の中、OpenAIが提唱したGDPvalは、AIの評価軸を「抽象的な推論能力（Reasoning）」から「具体的な経済的アウトプット（Economic Output）」へと移行させる試みである。これはAI評価における「経済的転回（Economic Turn）」とも呼べる動きであり、AIを単なる情報処理システムとしてではなく、GDPを生み出す生産要素として再定義するものである[1]。

2.2 GDPvalの設計思想と目的

GDPvalの名称は「Gross Domestic Product valuable tasks」に由来し、その目的は「AIが経済的にどれだけの価値を生み出せるか」を定量化することにある[1]。OpenAIのチーフエコノミストであるRonnie Chatterji氏らが主導するこのプロジェクトは、AIの能力をビジネスリーダーや政策立案者が理解可能な言語（すなわち、時間短縮、コスト削減、品質維持）で表現することを目指している[11]。

この指標は、単にモデルの優劣を決めるランキングのためだけでなく、企業がAI導入のROIを試算するための基礎データを提供し、どの業務領域が自動化に適しているかを示唆するロードマップとしての機能も期待されている[2]。

---

3. GDPvalフレームワークの構造と方法論

3.1 9つの主要セクターと44の職業選定

GDPvalの最大の特徴は、評価対象となるタスクの選定プロセスにある。OpenAIは米国労働統計局（BLS）のデータを基に、米国GDPへの付加価値貢献度が5%を超える主要なセクターを特定し、そこからさらに雇用者数と賃金総額が大きい「知識労働」を中心とした44の職業（Occupations）を抽出した[1]。

表1：GDPvalが対象とするセクターと職業の内訳

セクター	GDP貢献度	選定された職業の例	経済的規模（賃金総額）
専門・科学・技術サービス	8.1%	ソフトウェア開発者、弁護士、会計士、プロジェクト管理スペシャリスト	ソフトウェア開発者だけで約2,390億ドル
医療・社会扶助	7.6%	正看護師、ナースプラクティショナー、医療サービス管理者	正看護師だけで約3,230億ドル

製造業	10.0%	機械エンジニア、工業エンジニア、購買代理人、生産管理監督者	生産管理監督者だけで約510億ドル
金融・保険	7.4%	金融アナリスト、財務マネージャー、個人ファイナンシャルアドバイザー	財務マネージャーだけで約1,470億ドル
政府・行政	11.3%	コンプライアンス担当官、行政サービス管理者、警察・探偵の監督者	コンプライアンス担当官だけで約330億ドル
情報	5.4%	プロデューサー・ディレクター、編集者、ジャーナリスト	プロデューサー等で約166億ドル
小売業	6.3%	薬剤師、小売販売監督者、一般業務マネージャー	一般業務マネージャーだけで約4,770億ドル
卸売業	5.8%	営業マネージャー、卸売販売代理人、受注係	営業マネージャーだけで約970億ドル
不動産・賃貸	13.8%	不動産ブローカー、資産管理者、カウンター係	資産管理者だけで約245億ドル

これらの職業群は、物理的な操作を伴わずにコンピュータ上で完結可能な業務（Digital Knowledge Work）に焦点が当てられている。これは現段階のAIモデル（LLM）がロボティクスと統合されていないことを前提とした現実的な区分であるが、同時に後述する「物理的タスクの除外」というバイアスも内包している[12]。

3.2 タスク生成プロセス：「ゴールド・デリバラブル」の構築

GDPvalにおける「タスク」は、AI研究者が作成した合成データではなく、選定された44の職業に従事する平均14年の実務経験を持つ業界の専門家によって作成されている[2]。

- タスクのリアリティ：タスクは、専門家が日常的に行っている業務そのものを反映しており、例えば「バイオテック企業における変更管理SOP（標準作業手順書）のドラフト作成」や「監査のためのサンプリング計画の策定」などが含まれる[15]。
- コンテキストの深さ：従来のベンチマークが短いテキストプロンプトであったのに対し、GDPvalのタスクには、業務遂行に必要な背景情報として、スプレッドシート、PDF契約書、CAD図面、音声ファイル、過去のメール履歴など、最大で数十個の参照ファイルが添付される[14]。
- ゴールド・デリバラブル（模範解答）：タスクを作成した専門家自身が、そのタスクに対する自身の解答（成果物）を作成する。これが「ゴールド・デリバラブル」と呼ばれ、AIの出力を評価する際の品質基準（ベースライン）となる[1]。

3.3 評価メカニズム：ブラインドペアワイズ比較

GDPvalの採点方法は、正解との完全一致（Exact Match）を見るような自動化された手法ではなく、人間の主観的判断を重視した定性評価を採用している。

1. ブラインドテスト：評価者（同職種の別の専門家）には、「人間の専門家が作成した成果物（ゴールド）」と「AIモデルが生成した成果物」の2つが、作成者が伏せられた状態で提示される。
2. 多角的評価基準：評価者は、単なる事実の正確性だけでなく、構成の論理性、文体の適切さ、フォーマットの美しさ、そして「実務でそのまま使えるか（Usability）」といった観点から総合的に比較を行う[3]。
3. 評価コスト：1つのタスク比較に専門家は平均して1時間以上を費やす。これは評価プロセスの信頼性を高める一方で、大規模な実施を困難にする高コスト要因となっている[8]。

---

4. 最新モデルによる評価結果の分析

4.1 GPT-5.2 Thinkingのパフォーマンス：70%の壁

OpenAIが2025年後半に発表したGPT-5.2シリーズ（特にThinkingモデル）に関する初期結果は、AIが「明確に定義された知識労働」において人間を凌駕しつつあることを示唆している。

表2：主要モデルのGDPval勝率比較

モデル	勝率（Win）+ 引き分け（Tie）	特記事項
GPT-5.2 Pro	**74.1%**	最上位モデル。推論能力とツール使用の最適化が顕著[5]。
GPT-5.2 Thinking	**70.9%**	人間の専門家に対し7割以上の確率で同等以上の成果物を生成[4]。

Claude Opus 4.1	~49.0%	美的感覚やフォーマット維持に強みを持つが、GPT-5系には及ばず[2]。
GPT-5.1 Thinking	38.8%	前世代モデルからの飛躍的な向上（約2倍のスコア）を確認[5]。

この「70.9%」という数字は、AIが補助ツールを超えて、特定のタスクにおいては「シニアレベルの専門家」と同等の品質を提供できるようになったことを意味する。特に、表計算ソフトの操作や複雑なドキュメント作成において高いスコアを記録している[4]。

4.2 圧倒的な経済効率性：「100倍」の衝撃

GDPvalの結果から導き出される最も強力な経済的メッセージは、品質そのものよりも、その品質を達成するためのコストと時間の圧倒的な差にある。

- 時間短縮：人間の専門家が完了までに平均7〜9時間を要するタスクを、AIモデルは数分から数十分で完了させる。これはおよそ11倍から100倍のスピードアップに相当する[2]。
- コスト削減：専門家に支払う時給（例えば弁護士やエンジニアの高額な報酬）と比較して、AIのAPI利用料や推論コストは100分の1以下であると試算されている[2]。

このデータは、Databricksなどの企業が顧客に対してAI導入を推進する際の強力な論拠となっており、「人間が行うよりも100倍安く、同等の品質のドラフトを作成できる」という事実は、企業のコスト構造を根本から変える可能性を秘めている[2]。

4.3 失敗の質の変化

AIが人間に敗北したケース（約30%）の内訳分析も興味深い洞察を提供している。

- 許容範囲内の劣後（**Acceptable but subpar**）：敗北の約47%は、成果物として致命的な誤りはないものの、人間の専門家の方がより洗練されていたり、文脈を深く汲み取っていたりするケースであった[8]。
- 致命的な失敗（**Catastrophic failure**）：事実の捏造や重大な計算ミス、指示の完全な無視といった「実務で使うと危険な失敗」は、全体の約3%にとどまっている。この「3%」を許容できるかどうかが、実務適用の最大の障壁となる[8]。

---

5. 比較検証：GDPvalとRemote Labor Index (RLI) の乖離

GDPvalが示す「AIは人間並みである」という楽観的な結論に対し、強力なカウンターパートとして存在するのが、Scale AIやCenter for AI Safetyなどが発表した「Remote Labor Index (RLI)」である。両者は共に「現実の仕事」を対象としているが、その結果には衝撃的な乖離が存在する。

5.1 70% vs 2.5%：埋まらない溝

- **GDPval**: 人間の専門家に対する勝率（Win+Tie）が約**70%**。
- **Remote Labor Index (RLI)**: AIが完全に自動化できたタスクの割合は**約2.5%**未満[6]。

この極端な差は、両ベンチマークが想定する「仕事」の定義の違いに起因している。

表3：GDPvalとRLIの構造的比較

特徴	GDPval (OpenAI)	Remote Labor Index (RLI)
タスクの定義	明確に仕様化された知識労働 (Well-specified)	曖昧で複雑なフリーランス案件 (Real-world freelance)
入力情報	整理されたプロンプトと参照ファイル	Upwork等の実際の案件概要（ブリーフ）、未整理のファイル群
成果物の形式	主に文書、表計算、プレゼン資料	3Dモデル、CAD、動画編集、完全なアプリ、複雑なコード
完了条件	専門家による相対評価で「良い」とされるか	実際のクライアントが「納品」を受け入れるか（実用性）
主な失敗要因	指示の不履行、ニュアンスの欠如	ファイル破損、フォーマット不整合、一貫性の欠如、要件の誤解

5.2「明確な指示」と「曖昧な現実」

GDPvalのタスクは、専門家によって「AIに解かせるために」ある程度整理された状態で提示される。プロンプトには必要な制約条件や期待される出力形式が明記されている（Well-specified）。一方、RLIは実際のフリーランス市場（Upworkなど）から取得された「生の案件」を使用する。ここでの指示は曖昧であり、「いい感じのロゴを作ってくれ」「この動画を編集して」といった、文脈理解と自律的な判断（エージェンシー）を強く要求するものが多い[6]。

RLIの結果は、現在のAIが「指示された通りの文章を書く」ことには長けているが、「曖昧な依頼から意図を汲み取り、複数のツールを駆使してプロジェクトを完遂する」能力には依然として欠けていることを示している。特に、3Dモデリングや動画編集といったマルチモーダルな出力において、AIはファイル形式を破損させたり、一貫性のない成果物を出したりする傾向が強い[19]。

5.3 結論の相対化

この比較から導き出される結論は、**「GDPvalはAIが得意な領域（テキスト・論理処理・定型業務）に焦点を当てており、RLIはAIが苦手な領域（エンドツーエンドの自律遂行・非定型マルチメディア）を含んでいる」**ということである。GDPvalの結果のみを見て「AIはすべての仕事を代替できる」と考えるのは危険であり、RLIの結果のみを見て「AIは役に立たない」と考えるのもまた早計である。真実はその中間にあり、タスクの性質によってAIの有用性は0%から100%の間で劇的に変動する。

---

6. 批判的分析：方法論的欠陥と構造的課題

GDPvalに対しては、RLIとの比較以外にも、その方法論や前提に対する鋭い批判が学術界や産業界から寄せられている。

6.1「66%問題」：評価インフラのボトルネック

Pranil Dasikaらによる分析は、GDPvalが直面している最も深刻な工学的課題を指摘している。それは「AIの性能」ではなく「評価の信頼性」の問題である[8]。

- 自動評価の限界：OpenAIは評価コストを下げるためにGPT-5ベースの自動評価モデルを開発したが、このモデルの人間の評価者との一致率は**66%**にとどまった。
- 人間の不一致：さらに衝撃的なのは、人間の専門家同士の評価一致率も**71%**に過ぎないという事実である。つまり、人間の専門家同士でも約3割の確率で「何が良い仕事か」について意見が食い違う。
- 品質保証の壁：評価基準がこれほど不安定である場合、企業がAIシステムを本番環境に導入する際、「品質保証（QA）」を自動化することが極めて困難になる。66%の精度しかない「検品機」では、残りの34%の判定には依然として人間が介入しなければならず、AIによる「完全自動化」の夢は、この「評価の人間依存」によって阻まれることになる[8]。

6.2 ワンショット・バイアスと「プロンプト最適化の罠」

現実の知識労働は、一度の指示で完璧な成果物が出ることは稀である。通常は、上司やクライアントとの対話、ドラフトの修正、要件の再定義といった「反復プロセス（Iterative Process）」を経て完成する。

- 静的な評価：GDPvalは基本的に「ワンショット（一回のプロンプトと参照ファイル）」でタスクを完了させることを求めている。これにより、AIが本来得意とする「対話による明確化」や「フィードバックループ」が評価から除外されている[1]。
- 要件定義の省略：現実の仕事で最も価値があり、かつ難しいのは「何を作るべきか」を定義すること（要件定義）だが、GDPvalではこの部分がプロンプトとして既に与えられているため、AIは「仕様書通りの実装」を行うだけでよい。これは業務の難易度を実質的に下げている可能性がある[20]。
- プロンプト最適化の罠：企業がこのベンチマークを見て「プロンプトさえうまく書けばAIは使える」と誤解するリスクがある。実際には、プロンプトエンジニアリングよりも、システム全体の設計や

フィードバックループの構築（Evaluation Engineering）の方が重要である[8]。

6.3 経済的還元主義（GDP Fallacy）とバイアス

経済学者や社会科学者からは、GDPvalが「経済的価値（GDP）」を「仕事の価値」と同一視している点に対する懸念が表明されている[10]。

- 複雑性の無視：GDPは市場取引された価値のみを測定するため、組織内でのメンタリング、信頼構築、リスク管理、調整業務といった「価格のつかない労働（Non-market Labor）」が評価から抜け落ちる。
- 関係性労働の軽視：看護師や教師などの職業において、成果物の作成（ケアプラン等）は業務の一部に過ぎず、対人関係や感情労働（Emotional Labor）が重要な割合を占める。GDPvalはこれらをデジタル成果物のみに還元して評価しており、AIがこれらの職業全体を代替可能であるという誤った認識を与えるリスクがある。
- 米国中心主義：GDPvalは米国労働統計局のデータに基づいているため、他国の労働慣行や法規制、文化的な「良い仕事」の定義とは必ずしも一致しない可能性がある[10]。

---

7. セクター別の詳細分析と影響予測

GDPvalの結果と批判を踏まえ、主要セクターごとのAI導入の影響を分析する。

7.1 専門・技術サービス（法務、会計、コンサルティング）

- 適合度：高。契約書レビュー、監査計画、財務分析などは「正解」や「フォーマット」が明確であり、GDPvalのスコアが現実の生産性に直結しやすい。
- 課題：責任の所在と「幻覚（Hallucination）」リスク。3%の致命的失敗が許されない領域であるため、AIはドラフト作成に留まり、人間のレビューが必須となる[21]。

7.2 医療・社会扶助

- 適合度：中～低。診断書の要約やシフト管理などの事務タスクでは有効だが、GDPvalが測定しているのはあくまで「デジタルタスク」である。患者との対話や身体的ケアといった本質的業務には適用できない。
- リスク：RLIとの乖離が示すように、曖昧な状況判断（トリアージ等）においてAIはまだ信頼性に欠ける[21]。

7.3 クリエイティブ・情報産業（メディア、デザイン）

- 適合度：中。記事の執筆や要約は得意だが、RLIの結果が示す通り、3D制作や複雑な動画編集などのマルチモーダルタスクでは、ファイルの整合性維持に課題がある。
- 変化：クリエイターの役割は「素材を一から作る」ことから「AIが生成した大量の候補から選び、修正する（ディレクション）」ことへとシフトする。

---

8. 労働市場と企業戦略への示唆

8.1 職業の変容：代替ではなく「タスク分解」

GDPvalの結果は、44の職業がそのままAIに置き換わることを意味しない。むしろ、職業を構成する「タスク」のレベルでの分解（Unbundling）が進むことを示唆している[1]。

- コモディティ化するスキル：レポート作成、データ集計、基本的な翻訳などのタスクは、AIによって極めて低い限界費用で実行可能になるため、これらのスキル自体の市場価値は低下する。
- 価値の移動：人間にとっての価値は、「成果物を作る」ことから、「成果物の品質を定義・評価する（目利き）」、「AIに適切なコンテキストを与える（指示）」、「AIがカバーできない対人・物理タスクを担う」ことへとシフトする[11]。

8.2 企業における**AI**導入戦略の変化：**ROI**重視へ

GDPvalのようなベンチマークの登場により、企業のAI導入戦略は「実験（PoC）」から「ROIに基づく実装」へと移行しつつある[18]。

- 「人間にしかできない**34%**」への投資：評価の自動化が完全にはできない（66%の一致率）以上、最終的な品質責任を持つ人材の重要性は逆に高まる。企業は「AIオペレーター」ではなく、AIの出力を批判的に評価できる「AIスーパーバイザー」を育成する必要がある。
- 「失敗」のマネジメント：AI導入の成否は、約30%の確率で発生する「人間以下の出力」をいかに効率的に検知し、修正するプロセス（Human-in-the-loop）を構築できるかにかかっている[8]。

8.3 5段階の**AGI**レベルと**GDPval**の位置づけ

OpenAIなどはAGI（汎用人工知能）への到達度を5段階で定義しているが、GDPvalはその進捗を測る重要なマイルストーンとなる[23]。

- レベル**2**（推論者）の証明：GDPvalにおける高い勝率は、AIが「レベル2：人間と同等の推論能力を持つ」段階に到達しつつあることを強力に裏付けている。
- レベル**3**（エージェント）への壁：しかし、RLIにおける低い成功率は、AIが「レベル3：自律的にタスクを遂行するエージェント」にはまだ達していないことを示している。今後のAI開発の焦点は、単発のタスク品質向上から、長期間にわたる自律的遂行能力の獲得へと移るだろう。

---

9. 結論：**AI**活用のための羅針盤として

OpenAIのGDPvalは、AIの評価を「実験室の知能テスト」から「実社会の経済活動」へと引き上げた点で、AI開発史における重要な転換点である。70.9%という高い勝率は、条件が整った環境下であれば、AIが既に熟練した専門家に匹敵する知的生産能力を持つことを実証している。

しかし、その「条件が整った環境（Well-specified tasks）」という前提こそが、現在のAI実装における最大の落とし穴である。現実の仕事の曖昧さ、複雑さ、そして評価の主観性は、GDPvalが想定するモデルよりも遥かに厄介である。RLIとの比較が示す通り、AIは「部分的なタスク」では超人的であっ

ても、「仕事全体」を任せるにはまだ未熟な存在である。

したがって、企業や政策立案者は、GDPvalの数値を「職業代替のカウントダウン」として受け取るべきではない。むしろ、「どのタスクが**AI**によってコモディティ化され、どのプロセスに人間が介入すべきか」を判断するための羅針盤として活用すべきである。AIは「安価で高速だが、時折指示を誤解し、品質管理を必要とする部下」であり、その部下を使いこなすためのマネジメント能力（評価・要件定義・統合）こそが、今後の経済的価値の源泉となるだろう。

免責事項: 本報告書は、2025年12月時点での公開情報および調査資料に基づいて作成されており、AI技術の急速な進展により、状況は変化する可能性がある。

引用文献

1. Measuring the performance of our models on real-world tasks | OpenAI, 12月 12, 2025にアクセス、https://openai.com/index/gdpval/
2. OpenAI's GDPval Framework Sets New Standard for Measuring AI's ..., 12月 12, 2025にアクセス、https://mlq.ai/news/openais-gdpval-framework-sets-new-standard-for-measuring-ais-economic-impact/
3. gdpval: evaluating ai model performance - OpenAI, 12月 12, 2025にアクセス、https://cdn.openai.com/pdf/d5eb7428-c4e9-4a33-bd86-86dd4bcf12ce/GDPval.pdf
4. How GPT-5.2 stacks up against Gemini 3.0 and Claude Opus 4.5, 12月 12, 2025にアクセス、https://www.rdworldonline.com/how-gpt-5-2-stacks-up-against-gemini-3-0-and-claude-opus-4-5/
5. Introducing GPT-5.2 - OpenAI, 12月 12, 2025にアクセス、https://openai.com/index/introducing-gpt-5-2/
6. HOT TAKE: Can AI do 2.5% of your work or 48% of your work?, 12月 12, 2025にアクセス、https://www.aisidequest.com/p/hot-take-can-ai-do-2-5-of-your-work-or-48-of-your-work
7. Measuring AI Automation of Remote Work - Remote Labor Index, 12月 12, 2025にアクセス、https://www.remotelabor.ai/paper.pdf
8. OpenAI's GDPval: Why the 66% Automated Grading Problem ..., 12月 12, 2025にアクセス、https://medium.com/@pranil.dasika/openais-gdpval-why-the-66-automated-grading-problem-matters-more-than-the-48-win-rate-a5e542508196
9. OpenAI's New Benchmark Shows AI Does Knowledge Work 100X ..., 12月 12, 2025にアクセス、https://www.marketingaiinstitute.com/blog/openai-gdpval
10. (PDF) Critical Analysis: The Fundamental Flaws in OpenAI's GDPval ..., 12月 12, 2025にアクセス、https://www.researchgate.net/publication/395874739_Critical_Analysis_The_Fundamental_Flaws_in_OpenAI's_GDPval_Evaluation_Framework

11. OpenAI launches workforce blueprint for AI skill preparedness, 12月 12, 2025にアクセス、https://www.upskillist.com/blog/openai-workforce-blueprint-ai-skill-preparedness/
12. GDPval: Evaluating AI Model Performance on Real-World ... - arXiv, 12月 12, 2025にアクセス、https://arxiv.org/html/2510.04374v1
13. GDPval: Evaluating AI Model Performance on Real-World ... - arXiv, 12月 12, 2025にアクセス、https://arxiv.org/html/2510.04374
14. Will AI Replace My Job In 2025? Definitive, Honest GDPval Insight, 12月 12, 2025にアクセス、https://binaryverseai.com/will-ai-replace-my-job-openai-gdpval/
15. OpenAI's GDPval Explained | IPM - Institute of Project Management, 12月 12, 2025にアクセス、https://instituteprojectmanagement.com/blog/openais-gdpval-explained/
16. openai/gdpval · Datasets at Hugging Face, 12月 12, 2025にアクセス、https://huggingface.co/datasets/openai/gdpval
17. GPT-5.2 Review: 70% GDPval Score Crushes Industry Experts, 12月 12, 2025にアクセス、https://binaryverseai.com/gpt-5-2-review-70-gdpval-score-benchmarks-price/
18. AI models now match human expert performance on business tasks ..., 12月 12, 2025にアクセス、https://inkeep.com/blog/gdpval-ai-expert-performance
19. Remote Labor Index: Top AI Agents Successfully Complete 2.5% of ..., 12月 12, 2025にアクセス、https://neurohive.io/en/news/remote-labor-index-top-ai-agents-successfully-complete-2-5-of-freelance-projects/
20. New benchmark for economically viable tasks across 44 ... - Reddit, 12月 12, 2025にアクセス、https://www.reddit.com/r/singularity/comments/1nqef1l/new_benchmark_for_economically_viable_tasks/
21. OpenAI Benchmark Tests AI Productivity as CFOs Demand ROI, 12月 12, 2025にアクセス、https://www.pymnts.com/news/artificial-intelligence/2025/openai-benchmark-puts-ai-productivity-to-the-test-as-cfos-demand-roi/
22. UpBench: A Dynamically Evolving Real-World Labor-Market Agentic ..., 12月 12, 2025にアクセス、https://www.upwork.com/static/webflow/assets/webflow-human-agent-productivity-index/upbench_paper.pdf
23. The 5 Levels in Achieving AGI – Genesis, 12月 12, 2025にアクセス、https://genesishumanexperience.com/2024/12/22/the-5-levels-in-achieving-agi/
24. If You Don't Understand Level 2 AI, You'll Be Left Behind, 12月 12, 2025にアクセス、https://aakashgupta.medium.com/were-only-on-step-2-of-5-to-agi-but-most-teams-are-building-for-the-wrong-level-2e4df4f2e90e